A ROUGH GUIDE TO THE ACOUSTIC-TO-ARTICULATORY INVERSION OF SPEECH

Asterios Toutios, Konstantinos Margaritis

Abstract— **This article reviews a specific speech research area called acoustic-to-articulatory inversion of speech, or speech inversion, which refers to the problem of mapping the acoustic speech signal onto a space describing the configuration of the human vocal tract that actually produced this signal. This space may be modeled in a variety of ways, such as with trajectories of the movement of the articulators - certain parts of the human vocal tract- derived by means of some specialized medical imaging device, or by using linguistics-related abstract classes to describe the evolution of the articulatory state through time. The problem is by far non-trivial, mainly due to the one-to-many nature and the high non-linearity of the acoustic to articulatory mapping. Numerous approaches towards its solution have been proposed. Such a solution could have several applications, with most important the probability of improving the performance of current automatic speech recognition systems.**

Keywords— **Acoustic-to-Articulatory Inversion, Detection of Phonological Features, Speech Recognition, Machine Learning**

I. Introduction

HUMAN speech is, from a strictly mechanical point of view, the effect of the air flowing from the lungs to the acoustic environment, through the human vocal tract. This airflow is normally constricted by the various parts of the human vocal tract, such as the vocal cords, the tongue, the palate, the teeth or the lips. The variety of the levels and types of this constriction is the actual cause of the variety of the sounds that a human can produce and that constitute the units forming speech, perhaps the greatest expression of human intelligence.

The parts of the human vocal tract, such as the forementioned ones, which have a role in the production of speech, are called the *articulators*. The positioning of the body of them, which is obviously of crucial importance for speech production, is called the *articulatory state*.

Still from the same mechanical viewpoint, the final manifestation of human speech is the acoustic signal. Under normal conditions, humans are quite accustomed with it, since they can perceive it with the sense of hearing. It may be recorded with readily available devices and quite numerous techniques have been developed in order to process and analyze it.

Apparently, the articulatory state and the corresponding acoustic signal bear a strong relationship among each other. We may view two spaces, an acoustic space and an articulatory one as well as a process of mapping each one onto the other. If the articulatory state is known in some detail, the acoustic signal can easily be derived. This is an actual physical process - the human speech production mechanism - which is quite straightforward.

What is of concern here is the inverse case. Namely, if the acoustic signal is known (something which is the usual case) can the articulatory state be somehow calculated or even approximated?

The recovery of the articulatory state given the acoustic signal is not a trivial problem. Not having a direct analytical solution, it is considered a difficult and ill-posed problem, puzzling researchers for over three decades now, and being given the status of becoming a whole research area, called *acoustic-to-articulatory inversion of speech* or, more simply, *speech inversion*. Of course there are reasons for this difficulty.

One of the first things one has to consider is the way the articulatory state will be described, or modeled. It is a rather complicated problem to which we will refer later in this article. Another factor that contributes to making the speech inversion problem hard to solve, is the "one-to-many" nature of the acoustic-to-articulatory mapping. A given articulatory state has always only one acoustic realization. But, from the other side, a given acoustic signal may be the outcome of more than one articulatory states. Furthermore, the mapping is highly non-linear. A slight variation of the articulatory state may give rise to a whole different acoustic signal. To picture this one may think of the extreme case of the ventrilloquist, where the articulators seem to be static, while a plethora of sounds are being heard.

The motivation behind the ongoing research on acoustic-to-articulatory inversion despite the inherent difficulties of the problem seems to arise from the potential applications of a succesful solution. Perhaps the most interesting one is the possibility of using the additional articulatory information derived from such a solution in order to improve the performance of current automatic speech recognition systems, especially in cases such as with noisy, spontaneous or pathological speech. Such a possibility is demonstrated in several recent papers, where the articulatory information is embedded in speech recognition systems by various means, such as Bayesian Networks or factorial Hidden Markov Models. Other proposed applications include speech synthesis, building visual aids for teaching hearing

The authors are with the Parallel and Distributed Processing Laboratory, Department of Applied Informatics, University of Macedonia, 156 Egnatia Str., P.O. Box 1591, 54006, Thessaloniki, Greece. E-mail: {*toutios*, *kmarg*}*@uom.gr*.

impaired people how to speak and as a means of study in phonetics and phonology.

Another important expected outcome from the study of the acoustic-to-articulatory inversion problem is the modeling of coarticulation. *Coarticulation* is a term describing the manner the acoustic manifestation of a particular phoneme is dependent by its context. In other words that a phoneme may sound differently depending on the phonemes it is surrounded by. Classic approaches to automatic speech recognition deal with coarticulation by considering biphones or triphones instead of phonemes as atomic speech units, or classes, to be modeled, thus increasing by far the number of these classes, and needing far more training examples.

II. Manifestations of the Articulatory Space

We have already mentioned that an important first thing to consider when dealing with the acoustic-to-articulatory inversion problem is the manner by which the articulatory state will be described. There is a number of possible suggestions.

Various theoretical models have been developed in the past in order to describe the shape of the human vocal tract during speech production. They have been used extensively in early works on speech inversion and still do in some degree. The science of linguistics, and particularly phonetics, offers a second broad class of models for describing the articulatory state, though in a somewhat abstract fashion. Some writers refer to works using these models with the term "Detection of Phonological Features" instead of "Acoustic-to-Articulatory Inversion" but, in this article, we will prefer a unified view of those. Finally, the most promising manner of describing the articulatory state is related to the use of some specialized medical imaging devices that actually track and record the movement of the human articulators during the production of speech.

A. Theoretical Models

Several theoretical models that describe the human speech production process have been proposed until now. We will consider here two of the most important ones, in the context of the acoustic-to-articulatory inversion field, namely Maeda's model [1] and the lossless tube model [2].

A.1 Maeda's Model

Maeda's articulatory model represent the vocal tract geometry with seven articulatory parameters: three for the tongue (tongue dorsum, tongue body and tongue tip), two for the lips (opening and protrusion), one parameter for the jaw and one for the larynx.

Maeda's model needs to be adapted before being applied to a particular speaker [3]. In Figure1 a visualization of Maeda's model adapted for a particular speaker is shown.

A.2 The Lossless Tube Model

The human vocal tract is simulated by a straight tube through which air is blown. It has been found that a tube with a curvature does not sound much different than from one without. Furthermore, it is assumed that the tube is lossless. This means that energy loss due to friction between the flowing air and the tract walls is not modeled.

The model can be extended further so that a series of connected tubes are modelling the constrictions in the vocal tract. A set of parameters describing this model can be obtained by figuring out where air pressure waves hit the walls and get reflected. This phenomenon can be quantified as entities known as reflection coefficients for each tube [4].

In figure 2 the lossless tube model is illustrated.

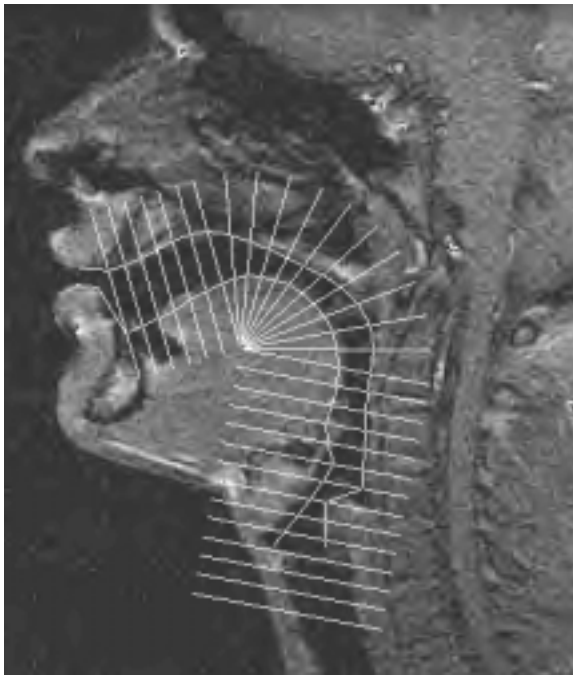

Figure 1. White lines visualize Maeda's model adapted for a particular speaker (Figure taken from http://www.loria.fr/)

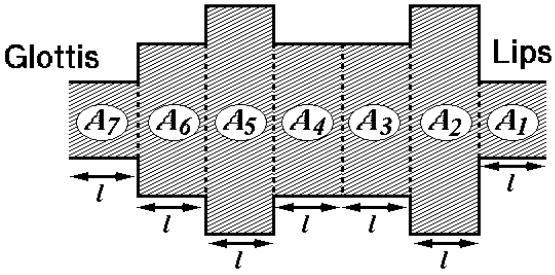

Figure 2. A simple illustration of the lossless tube model (Figure taken from http://isl.ira.uka.de/)

B. Linguistics-Derived Models

Using knowledge of linguistics, and particularly phonetics, any given phoneme of a spoken language may be related to a vector of features that describe by qualitative means the corresponding articulatory state. Sometimes these features may also have a functional, as opposed to a srtictly articulatory, meaning.

Phonetic features have also been proposed as the basis of spoken language universals, in the sense that while

Table I

Example of a multi-valued feature system

Features	Values
voicing	voiced, voiceless, silence
manner	vowel, nasal, lateral, approximant, fricative, silence, stop
place	dental, coronal, labial, retroflex,velar, glottal, high, mid, low, silence
front-back	front, back, nil, silence
rounding	rounded, unrounded, nil, silence

Table II

Conversion of phones into multi-valued features

Phones	Features
[aa]	voiced, vowel, low, back, unrounded
[ow]	voiced, vowel, mid, back, rounded
[p]	unvoiced, stop, labial, nil, nil
[n]	voiced, nasal, coronal, nil, nil
[f]	unvoiced, fricative, labial, nil, nil

phonemes of a language vary, the set of features does not and is the same for all languages.

We may consider three broad classes of such features that, in a way, relate to periods in "the history of thought" for modern phonetics. These are the multi-valued features describing primarily the place and manner of articulation, the binary - or distinctive - features, and the Government Phonology primes.

B.1 Multi-Valued Phonetic Features

The articulatory state may be represented by means of abstract classes describing the most essential articulatory properties of speech sounds [5], [6]. Examples of such properties are *voiced, nasal, rounded* etc.

Several articulatory components, or dimensions, which are partially independent of each other interact in order to produce human speech. Examples of such dimensions are *voicing* which describes the state of the glottis and the activity of the vocal cords, the *manner of articulation*, i.e. the shape of a constriction made by an articulator in the vocal tract, and the *place of articulation*, which describes the actual location of the previous constriction. These components may take multiple values.

A variety of choices for the forementioned components or the values they may take can be found in the literature [7], [8]. In Table I the feature system selected in [8] is illustrated. Table II shows how some phones are decomposed into these features.

Table III

Conversion of phones into distinctive features

phone	[aa]	[ow]	[p]	[n]	[f]
vocalic	+	+	-	-	-
consonantal	-	-	+	+	+
high	-	-	+	-	-
back	+	+	-	-	-
low	+	-	-	-	-
anterior	-	-	+	+	+
coronal	-	-	-	-	-
round	-	+	-	-	-
tense	+	+	-	-	-
voice	+	+	-	+	-
continuant	+	+	-	-	+
nasal	-	-	-	+	-
strident	-	-	-	-	+

B.2 Distinctive Features

The principle of distinctive features was first proposed in the classic work of **Jakobson, Fant and Halle** in [9]. Although this work gained much attention when published, many regarded features as nothing more than a useful classification scheme, whereby one could refer to the class of "nasal phones" or "voiced phones". The power of features became evident with the publication of *The Sound Pattern of English* [10] where **Chomsky and Halle** showed that what were otherwise complex phonological rules could be written concisely if features where used rather than phones.

The feature system in *The Sound Pattern of English* uses production-based binary features. In this system, each phone is composed of a vector of 13 binary components which represent production features such as *voicing, high, low* (representing tongue position during vowels), *round* (for lip rounding), *continuant* (to distinguish continuous sounds such as vowels and fricatives from stops), and so on. A "+" or a "-" is indicating the presence or absence of a feature in a given phone. In Table III some examples of analysis of phones into these distinctive features are shown [7].

Slight variations of the previous system have also been proposed in the literature (e.g. [11]).

B.3 Government Phonology Primes

In *Government Phonology* [12] sounds are described by combining primes in a structured way, and phonological phenomena are accounted for by the fusing and splitting of primes within a sound.

The primes **A, I, U** and **@** are known as the *resonance primes*, and capture consonant and vowel sounds. They are derived from examination of the spectral properties of vowels. The **?** prime is present in sounds with a closure or

Table IV
Government Phonology Primes

phone		[aa]	[ow]	[p]	[n]	[f]
primes	**A**	*	*		*	
	I					
	U		*	*		*
	@					
	?			*	*	
	h			*		*
	H			*		*
	N				*	
head	**a**	*				
	i					
	u		*			

any abrupt and sustained decrease in amplitude. Frication (acoustically evident as aperiodic energy) is indicated by the presence of the **h** prime, and the nasal prime **N** is present in sounds with an articulatory oral closure. The **H** prime indicates unvoiced sounds, where the vocal folds are stiff and not vibrating periodically.

The vowels [a], [i], [u], [@] are represented by just a single prime while all other sounds are made by fusing primes. For example, fusing **A** and **U** gives [o] and fusing **A** and **I** produces [e]. More complex sounds, like diphthongs, require the primes to be arranged in a structured way. As well as simply fusing two or more primes, one of the primes can optionally be made the *head* of the expression, denoting its greater significance both phonologically and in determining the phonetic realisation of the sound.

In Table IV examples of government phonology primes for some phones are shown.

C. Medical Imaging Models

Perhaps the most interesting, in the context of the acoustic-to-articulatory inversion problem, way of modeling the articulatory space, involves the use of specialized medical imaging devices that draw information about the positioning or the activity of the articulators directly from the human subject during the process of speech production. The acquisition of such data is a very difficult and expensive task, nevertheless a few databases have been developed and are being made available, giving a boost to the relevant research.

C.1 Types of Medical Imaging Data

A number of techniques are used in order to partially acquire the articulatory state during speech directly from the human subject.

X-ray cineradiography [13] involves x-ray filming of the vocal tract during speech production. This particular technique is no longer used because of the danger of radiation

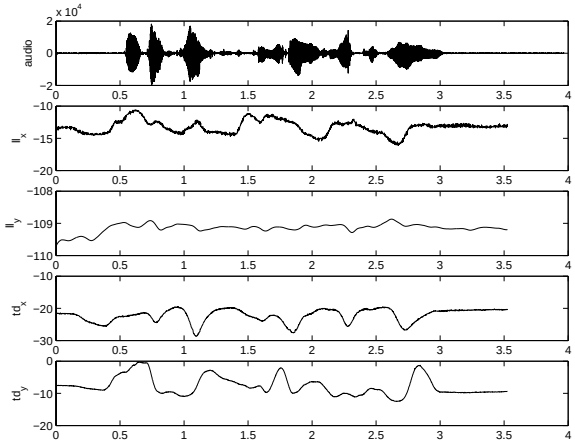

Figure 3. Example of EMA data. A female subject is uttering the phrase "We got drenched from the uninterrupted rain". From top to bottom the audio signal, the x-axis and y-axis trajectories for the lower lip and the x-axis and y-axis trajectories for the tongue dorsum are shown.

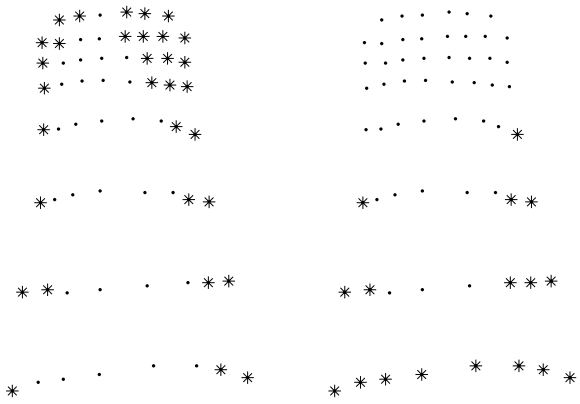

Figure 4. Example of EPG data. The tongue shape is apparent. Asterisks indicate the tongue-palate contact points.

overexposure. However, a lot of old x-ray films have been preserved and are available for research purposes.

For the acquisition of *Electromagnetic Misdagittal Articulography (EMMA)* or *Electromagnetic Articulography (EMA)* [14] data, sensor coils are attached to the human subject, on specific places on the lips, the teeth, the tongue and the velum. Then the human subject wears a special helmet which produces an alternating magnetic field that records the position of the coils at end points of small fixed-size time intervals. The outcomes are time-series, or trajectories, that illustrate the movement of the coils. Usually, there are two trajectories for each coil, one for the movement in the front-back direction of the head (x axis), and one for the top-bottom direction (y axis). An important characteristic of this trajectories is that they vary slowly in time. Figure 3 shows an example of EMA data.

In *electropalatography (EPG)* [15] the patterns of contact between the tongue and the palate during speech are determined. The technique utilises an artificial palate with 62 silver electrodes embedded in its tongue-facing surface. Figure 4 shows two instaces of electopalatographic data,

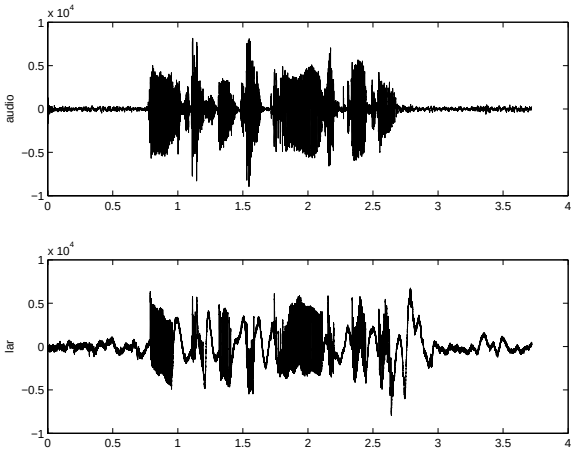

Figure 5. Audio and corresponding laryngographic signal of a male speaker uttering the phrase "My instructions desparetaly need updating"

where the asterisks indicate the points of the tongue that contact the palate.

Laryngography [16] or *electroglottography (EGG)* [17] involves the use of two electrodes which are placed on the throat of the speaker positioned on each side of the thyroid cartilage. A weak but constant voltage is passed from one electrode to the other allowing the current to fluctuate depending on the contact variations between the vocal cords. These variations are recorded giving the outcome that may be called the *laryngographic signal*. This signal includes high frequencies and resembles the actual audio speech signal. Figure 5 shows an example of a laryngogrphic signal.

Finally, *pneumotachography* [18] is a technique for measuring tha nasal and oral airflow velocity. Its actual use is for the diagnosis of respiratory problems such as asthma, but it might also be used as a source of additional information in the context of speech inversion.

C.2 Articulatory Databases

Three databases that include articulatory data, such as the ones described previously, are the MOCHA-TIMIT database, the EUR-ACCOR database and the X-Ray Wisconsin database.

The **MOCHA-TIMIT database** [19] is being developed by the Queen Mary University College in Edinburgh, Scotland. The goal was two gather data from 40 human subjects, but until the writing of this article data from only two of them, a male and a female, were freely available via ftp. The corpus, for each speaker, consists of 460 English sentences taken from the TIMIT continuous speech database [20] that are designed to include the main connected speech processes in English. Data include the audio signal sampled at 16 KHz, phonetic and orthographic transcripts, the laryngographic signal sampled at 16 KHz, EMA data from the lower and upper incisor, the lower and upper lip, the tongue tip, tongue blade, tongue dorsum and the velum, sampled at 500 Hz, EPG data sampled at 200 Hz, and video recordings of the front view of the mouth area

The **EUR-ACCOR database** [21] was developed as part of a European ESPRIT project. It contains data from seven European languages (Catalan, English, French, German, Irish Gaelic, Italian and Swedish). The corpus consists of a combination of nonsense items, real words and short sentences. There are 5-10 speakers per language. Gathered data include the audio signal sampled at 20 KHz, laryngographic signal sampled at 10 KHz, pneumotachography data sampled at 500 Hz, and EPG data sampled at 200 Hz.

Finally, the **X-ray Wisconsin database** [22] includes EMA-like data, gathered from the University of Wisconsin X-ray Microbeam facility.

III. Modeling the Acoustic Signal

One of the first things one has to consider in order to buid a speech recognition system is the choice of features by which the acoustic signal will be represented, as the raw signal would be too much for any such system to handle. This is called a front-end parametrization of the signal, and is something to consider also in the case of acoustic-to-articulatory inversion.

The Mel Frequency Cepstral Analysis of the speech signal [23] provides such a set of fetures, called the Mel Frequency Cepstral Coefficients, or MFCCs. The MFCCs are robust, contain much information about the vocal tract configuration regardless the source of excitations, and may be used to represent all classes of speech sounds. They are a classic choice for automatic speech recognition systems and are used in the vast majority of the speech inversion implementations we are about to present subsequently.

Some works on acoustic-to-articulatory inversion of speech also use the Linear Predictive Coding (LPC) [24] coefficients for the front-end parametrization of the speech signal despite the fact that in the context of speech recognition their use is quite obsolete since they have in a large extend been replaced by the MFCCs.

Another set of features that might also be used in the same sense are the Perceptual Linear Predictive (PLP) [25] coefficients. The PLP coefficients capture the way humans perceive the speech signal with the sense of hearing.

Finally, **Hansen** in [26] presents a set of acoustic parameters which are immediately associated to specific phonetic and articulatory phenomena, suggesting the possibility of their use for speech inversion tasks.

IV. Approaches Towards A Solution

Various approaches have been proposed in the quest for an optimal solution to the acoustic-to-articulatory inversion problem. Following a taxonomy of those found in [27] we may view codebook approaches, neural network approaches, constrained optimization approaches, analytical approaches as well as stochastic modeling and statistical inference approaches. In the following, we present some recent examples of these approaches.

For a thorough discussion of older approaches to the speech inversion problem, the interested reader is referred to the work of **Schroeter and Sondhi** in [28]. We also

have to point out the work of **Dusan** in [29] where methods of applying phonetic and phonological constraints to the acoustic-to-articulatory inversion problem are reviewed and discussed upoon.

A. Codebooks

Some acoustic-to-articulatory inversion methods use codebook lookup procedures combined with optimization approaches in order to perform the inversion. The articulatory space is quantized and the corresponding acoustic features are synthesized to form a codebook of acoustic and articulatory vector pairs. The quality of the expected articulatory trajectories, which are the result of the inversion process, is highly dependent on the initial solutions given by the codebook. Thus, it is important that the codebook gives a good coverage of the articulatory space.

Ouni and Laprie in [30] represent the codebook as a hierarchy of hypercubes. Each hypercube represents a region of the articulatory space where the articulatory-to-acoustic mapping is linear. For each acoustic entry the whole codebook is searched for the relative articulatory parameters to be retrieved.

B. Neural Networks

In neural network approaches to the acoustic-to-articulatory inversion problem, the parameters of some neural networks are trained to get a non-linear continuous mapping between the articulatory parameters and the acoustic features. Approaches of this kind are most useful when the articulatory space is represented by means of linguistics-derived abstract classes.

King and Taylor in [7] use recurrent neural networks to perform feature detection on three phonological feature systems, namely distinctive features, multi-valued features and government phonology primes. Their networks perform well, with the average acccuracy for a single feature ranging from 86% to 93%.

Kirchhoff in [8] uses a set of multilayer perceptrons to map between MFCC parameters and the set of multivalued articulatory features of Table I. She achieves accuracy rates up to 95% depending on the feature in question.

C. Constrained Optimization

Prado et al. in [31] present a constrained optimization approach for estimating the articulatory state from the speech signal. The scheme they use concatenates a gradient search, which is accelerated using an algorithm inspired by the Fletcher-Reeves method, a classic non-linear optimization approach, and a linear successive approximation which assures convergence near the optimal articulatory vector. Constraints are imposed on the articulatory parameters to avoid physiologically impossible vocal tract configurations.

D. Analytical Methods

Laprie and Mathiew [32] present an example of an analytical approach to the speech inversion problem. They use Maeda's articulatory model and a variational calculus method. Their method includes inherent coarticulation constraints in the definition of an energy function to be minimized analytically.

E. Stochastic Modeling and Statistical Inference

Pehaps the most up-to-date and promising class of solutions to the speech inversion problem is the one that is based on stochastic modeling and statistical inference methods.

Richmond in [33] uses EMA data from the MOCHA-TIMIT database and calls upon a mixture density network to perform the acoustic-to-articulatory inversion. His investigation shows that the mixture density network is very well suited for delivering the required functionality for performing the inversion mapping.

King and Wrench in [34] use dynamical system modeling (Kalman filtering) with EMA data. The speech signal is parametrized by means of LPC analysis. One of their conclusions is that the underlying physical mechanism of speech production is sufficiently linear not to require non-linear models; however, the acoustic observations do not have a linear relationship to the articulatory parameters.

Ramsay in [35] takes a non-linear filtering approach. He outlines a stochastic framework for adapting an artificial model to real speech from acoustic measurements alone, using the Expectation Maximization (EM) algorithm [36] and showing that the solution of the problem in a maximum-likelihood sense relies on solving an associated state estimation problem to gather statistics from the measurement data.

Dusan and Deng in [37] use the EM algorithm, with the E-step accomplished by the Iterated Extended Kalman filtering and smoothing [38], to estimate the articulatory model parameters. They use EMA data and test their method only on vowel tokens.

Finally, **Carreira-Perpiñán and Renals** in [39] use EPG data from the EUR-ACCOR database and PLP parametrization of the speech signal. They present a latent variable approach to the acoustic-to-articulatory mapping. In latent variable modeling, the combined acoustic and articulatory data are assumed to have been generated by an underlying low-dimensional process. A parametric probabilistic model is estimated and mappings are derived from the respective conditional distributions.

V. Articulatory Information for Speech Recognition

Current automatic speech recognition systems [40], [41] typically use Hidden Models, Neural Networks or hybrid schemes in order to perform a mapping between the acoustic speech signal and the corresponding words or phonemes. A language model is used to retrieve the a priori probabilities of the appearance of these language units. Apart from this language model, the only input source of such systems is the acoustic signal, parametrized in some way.

Speech recognition systems based on this approach achieve nowadays quite satisfactory results when dealing with normal, structured and noise-free speech. However,

this is not the case with noisy, spontaneous or pathological systems. On the other side, it is widely accepted that these systems have reached a plateau in terms of performance [42], [43].

So, there is actually a question of finding novel approaches to the automatic speech recognition problem. One of these approaches uses articulatory information in order to enhance recognition. This information usually may not be readily available for everyday applications and has to be retreived from the acoustic speech signal by means of some form of acoustic-to-articulatory inversion. Some recent works explore this concept.

Kirchhoff in [8] uses multi-valued abstract articulatory features extracted from the speech signal by means of a set of multilayer perceptrons as a souce of information for the recognition of clear, reverberant and noisy speech. Mixture of experts systems corresponding to three different kinds of input sources are considered: using acoustic features alone, articulatory features alone, and both of them simultaneously. The latter system is the outcome of the combination of the other two by means of a product rule. The results indicate that using articulatory features instead of acoustic ones doesn't present with much of an improvement, as the results are somewhat similar. Nevertheless, the combined system exhibits a significant improvement, especially in the noisy speech case. As a matter of fact the improvement increases as the speech-to-noise ratio gets lower.

Frankel and King in [44] use EMA data and linear dynamical modeling. They consider acoustic and both real and simulated articulatory data for a simple phone classification task. They conclude, just like Kirchhoff, that the use of combined acoustic and articulatory data actually improves recognition performance.

Richardson et al. in [45] introduce a type of a Hiddden Markov Model where each state represents an articulatory configuration. The state transition matrix is governed by dynamic constraints on articulator motion. They call this scheme a Hidden-Articulator Markov Model, or HAMM. Their model by itself does not produce better word recognition results compared to a standard, acoustic-based HMM, however a combination of the two systems does. It is suggested that the articulatory system makes in general different mistakes than the acoustic one. This fact is actually beneficial for the recognition task.

Stephenson et al. in [46] building upon the groundwork done by **Zweig** in [47], investigate the use of Dynamic Bayesian Networks (DBNs) for incorporating articulatory data with acoustic data in automatic speech recognition. During training, the articulatory data, which are derived from the X-ray Wisconsin database, are introduced as variables to the DBN, which is expected, during testing, to be able to infer the distribution of the articulatory positions given the observed acoustics, thus accomplishing the acoustic-to-articulatory inversion task as a sub-product of the recognition task itself.

Finally **Sawhney and Wheeler** in [48] attempt to use knowledge of distinctive features in the context of recognizing dysarthric speech.

VI. Conclusion and Future Work

The recovery of the articulatory space from the acoustic signal poses an intriguing problem which has attracted, and keeps attracting, the interest of researchers worldwide. The problem is not a trivial one; on the contrary, it is considered difficult and ill-posed. It incorporates a number of scientific disciplines such as signal processing, machine learning, phonetics and medicine. It may have several forms, depending on the manner the articulatory space is described. Various attempts to solve it have been proposed, using a wide range of methods and techniques.

A succesful solution to the acoustic-to-articulatory inversion problem could have numerous applications, perhaps the most important of them being its use in the context af automatic speech recognition. Indeed, it has been proven that knowledge of the articulatory state can enhance the performance of speech recognition systems; a possibly essential improvement considering the current state of such systems. The role of the inversion in this context is to provide the articulatory data. Work is still in an early stage; a fully functional automatic speech recognition system that uses articulatory information is yet to be developed.

From our perspective, we consider the acoustic-to-articulatory inversion of speech, not only as a challenging machine learning problem, but as a suitable testbed for the application of various stochastic modeling methods as well. We have already built an Elman Neural Network that detects distinctive features from the speech signal. Development and testing were done on a small, non-standard corpus, with encouraging results. In the near future, we are planning on working with EMA data from the MOCHA database, considering the problem as more of a time series processing one. The problem is to map one time series (the speech signal) to a set of others (the EMA trajectories). We are thinking of using Neural Networks to this end [49].

In this article, we have tried to roughly present the acoustic-to-articulatory inversion field of research, outlining its basic concepts and reviewing some of the most current approaches to it. Surely, our discussion is by far a non-exhaustive one, since speech inversion is a quite large and lively field, with more than a hundred of published works having been counted so far.

References

[1] S. Maeda, "Un modèle articulatoire de la langue avec des componsantes linairés," in *Actes 10èmes Journées d' Etude sur la Parole*, Grenoble, 1979, pp. 152–162.

[2] John D. Markel and Jr. Augustine J. Gray, *Linear Prediction of Speech*, Springer Verlag, Berlin, 1976.

[3] Slim Ouni and Yves Laprie, "A study of the main constriction of the vocal tract for french vowel using an acoustic-to-articulatory inversion method," in *International Congress of Phonetic Sciences*, 2003.

[4] John R. Deller, John H.L. Hansen, and John G. Proakis, *Discrete Time Processing of Speech Signals*, IEEE Press, New York, 2000.

[5] Marina Nespor and Αγγελική Ράλλη, *Φωνολογία*, Εκδόσεις Πατάκη, 1999.

[6] Mark Aronoff and Janie-Rees Miller, Eds., *The Handbook of Linguistics*, Blackwell Publishers, 2000.

[7] Simon King and Paul Taylor, "Detection of phonological features

in continuous speech using neural networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 333–353, October 2000.

[8] Katrin Kirchhoff, *Robust Speech Recognition using Articulatory Information*, Ph.D. thesis, University of Bielefeld, June 1999.

[9] R. Jakobson, G. M. C. Fant, and M. Halle, *Preliminaries to Speech Analysis: the Distinctive Features and their Correlates*, MIT Press, 1952.

[10] Noam Chomsky and Morris Halle, *The Sound Pattern of English*, MIT Press, 1968.

[11] Ειρήνη Φιλιππάκη Warburton, *Εισαγωγή στη Θεωρητική Γλωσσολογία*, Νεφέλη, 1992.

[12] John Harris, *English Sound Structure*, Blackwell, 1994.

[13] Kevin G. Munhall, Eric Vatikiotis-Bateson, and Yohichi Tokhura, "X-ray film database for speech research," *Journal of the Acoustical Society of America*, vol. 98, no. 2, pp. 1222–1224, 1995.

[14] Jack Ryalls and Susan J. Behrens, *Introduction to Speech Science: From Basic Theories to Clinical Applications*, Allyn & Bacon, 2000.

[15] William J. Hardcastle, "The use of electropalatography in phonetic research," *Phonetica*, vol. 25, pp. 197–215, 1972.

[16] Ronald J. Baken, *Clinical Measurements of Speech and Voice*, Singular Publishing, 1996.

[17] Martin Rothenberg and James J. Mashie, "Monitoring vocal fold abduction through vocal fold contact area," *Journal of Speech and Hearing Research*, vol. 31, pp. 338–351, September 1988.

[18] W.J. Sullivan, G. M. Peters, and P.L Enright, "Pneumotachographs: Theory and clinical use," *Respiratory Care*, vol. 29, no. 7, pp. 736–749, 1984.

[19] Alan A. Wrench and William J. Hardcastle, "A multichannel articulatory database and its application for automatic speech recognition," in *5th Seminar on Speech Production: Models and Data*, Kloster Seeon, Bavaria, 2000, pp. 305–308.

[20] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallet, and Nancy L. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," Tech. Rep., National Institute of Standards and Technology, Gaithersburg, MD, 1993.

[21] Alain Marchal and William J. Hardcastle, "Instrumentation and database for the cross-language study of coarticulation," *Language and Speech*, vol. 36, no. 2,3, pp. 137–153, 1993.

[22] John R. Westbury, "X-ray microbeam speech production database user's handbook," Tech. Rep., University of Wisconsin, Madison, 1994.

[23] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in contonuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–166, 1980.

[24] Lawrence H. Rabiner and Ronald W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[25] Hynek Hermansky, "Perceptual Linear Predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[26] Anya Varnich Hansen, "Acoustic parameters optimised for recognition of phonetic features," in *Eurospeech 97*, Rhodes, Greece, 1997, pp. 397–400.

[27] Sacha Krstuloviç, *Speech Analysis with Production Constraints*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, 2001.

[28] Juergen Schroeter and Man Mohan Sondhi, "Techniques for estimating vocal-tract shapes from the speech signal," *IEEE Transactions Speech and Audio Processing*, vol. 2, no. 1, pp. 133–150, 1994.

[29] Sorin Dusan, "Methods for intergrating phonetic and phonological knowledge in speech inversion," in *International Conference on Speech, Signal and Image Processing ICSSIP2001*, Malta, 2001.

[30] Slim Ouni and Yves Laprie, "Improving acoustic-to-articulatory inversion by using hypercube codebooks," in *International Conference on Spoken Language Processing ICSLP2000*, Bejing, China, 2000.

[31] P. P. L. Prado, E. H. Shiva, and D. G. Childers, "Optimization of acoustic-to-articulatory mapping," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'92*, 1992, vol. 2, pp. 33–36.

[32] Yves Laprie and B. Mathiew, "A variational approach for estimating vocal tract shapes from the speech signal," in *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP'98*, 1998, pp. 929–932.

[33] Korin Richmond, "Mixture density networks, human articulatory data and acoustic-to-articulatory inversion of continuous speech," in *Workshop on Innovation in Speech Processing WISP2001*, 2001.

[34] Simon King and Alan Wrench, "Dynamical system modeling of articulator movement," in *International Conference on Phonetic Sciences ICPhS'99*, San Francisco, USA, 1999.

[35] Gordon Ramsay, "A non-linear filtering approach to stochastic training of the acoustic-to-articulatory mapping using the EM algorithm," in *International Conference on Spoken Language Processing ICSLP'96*, 1996.

[36] Jeff A. Bilmes, "A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," Tech. Rep., University of Berkeley, 1997.

[37] Sorin Dusan and Li Deng, "Recovering vocal tract shapes from MFCC parameters," in *International Conference on Spoken Language Processing ICSLP'98*, 1998.

[38] A. M. Jazwinsky, *Stochastic Processes and Filtering Theory*, Academic, New York, 1970.

[39] Miguel Á. Carreira-Perpi nán and Steve Renals, "A latent variable modeling approach to the acoustic-to-articulatory mapping problem," in *International Conference on Phonetic Sciences ICPhS'99*, San Francisco, USA, 1999, pp. 2013–2016.

[40] Lawrence R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recogntion," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[41] Ronald A. Cole, Joseph Mariani, Hans Uszkoreit, Annie Zaenen, and Victor Zue, Eds., *Survey of the State of the Art in Human Language Technology*, Center of Spoken Language Understanding, Carnegie Melon University, Pittsburgh, USA, 1996.

[42] Sadaoki Furui and Chin-Hui Lee, "Robust speech recognition - an overview," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1995.

[43] Richard P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, pp. 1–15, 1997.

[44] Joe Frankel and Simon King, "Speech recognition in the articulatory domain: Investigating an alternative to acoustic HMMs," in *Workshop on Innovation in Speech Processing WISP2001*, 2001.

[45] Matt Richardson, Jeff Bilmes, and Chris Diorio, "Hidden-Articulator Markov Models for speech recognition," in *ISCA IRTW Conference an Automatic Speech Recognition*, Paris, France, 2000.

[46] Todd A. Stephenson, Hervé Bourlard, Samy Bengio, and Andrew C. Morris, "Automatic speech recognition using dynamic bayesian networks with both acoustic and articulatory variables," in *International Conference on Spoken Language Processing ICSLP2000*, Bejing, China, 2000.

[47] Geoffrey Zweig, *Speech Recognition with Dynamic Bayesian Networks*, Ph.D. thesis, University of California, Berkeley, 1998.

[48] Nitin Sawhney and Sean Wheeler, "Using phonological context for improved recognition of dysarthric speech," Tech. Rep., MIT Media Lab, 1999.

[49] Georg Dorffner, "Neural networks for time series processing," *Neural Network World*, vol. 6, no. 4, pp. 447–468, 1996.